

# Predicting Protein-Protein Interactions from Protein Domains Using a Set Cover Approach

Chengbang Huang, Faruck Morcos, Simon P. Kanaan, Stefan Wuchty,  
Danny Z. Chen, and Jesús A. Izaguirre

**Abstract**—One goal of contemporary proteome research is the elucidation of cellular protein interactions. Based on currently available protein-protein interaction and domain data, we introduce a novel method, Maximum Specificity Set Cover (MSSC), for the prediction of protein-protein interactions. In our approach, we map the relationship between interactions of proteins and their corresponding domain architectures to a generalized weighted set cover problem. The application of a greedy algorithm provides sets of domain interactions which explain the presence of protein interactions to the largest degree of specificity. Utilizing domain and protein interaction data of *S. cerevisiae*, MSSC enables prediction of previously unknown protein interactions, links that are well supported by a high tendency of coexpression and functional homogeneity of the corresponding proteins. Focusing on concrete examples, we show that MSSC reliably predicts protein interactions in well-studied molecular systems, such as the 26S proteasome and RNA polymerase II of *S. cerevisiae*. We also show that the quality of the predictions is comparable to the Maximum Likelihood Estimation while MSSC is faster. This new algorithm and all data sets used are accessible through a Web portal at <http://ppi.cse.nd.edu>.

**Index Terms**—Computations on discrete structures, graph algorithms, bioinformatics (genome or protein) databases, biology, genetics.

## 1 INTRODUCTION

A goal of contemporary proteome research is the elucidation of the structure, interactions, and functions of proteins that constitute cells and organisms. Genomics has already produced an incredible quantity of molecular interaction data, contributing to maps of specific cellular networks. Indeed, large-scale attempts have unraveled the complex web of protein interactions in organisms such as *S. cerevisiae* [1], [2], [3], [4], [5], [6], [7] and *P. falciparum* [8]. Most recently, attention focused on the first protein interaction maps of complex multicellular organisms such as *C. elegans* [9], [10], *D. melanogaster* [11], and *H. sapiens* [12].

Although large-scale experimental attempts to uncover the complex webs of protein interactions in various organisms are still in progress, theoretical considerations focus on the prediction of potential protein interactions. Pioneering methods drew on the observation that interacting protein domains tend to combine into a fusion protein [13], [14]. Another approach focused on the observation that functionally linked proteins tend to be either preserved or eliminated in evolution. Proteins having matching phylogenetic profiles strongly tend to be functionally linked [15], [16].

Investigations of the spatial protein structure suggest that the fundamental unit is a domain. Independent of neighboring sequences, this region of a polypeptide chain folds into a distinct structure and mediates the protein's biological functionality. Comparing organisms over all three kingdoms of life, eukaryotes increasingly tend to have multidomain proteins, while the proteomes of bacteria or archaea mostly provide single domain proteins [17]. Such domain architectures govern interactions among proteins (Fig. 1), offering a framework for prediction models. Interaction domain pair profiles [18] assess the potential presence of a particular interaction by clustering protein domains, depending on sequence and connectivity similarities. References [19], [20], [21] figure protein interactions from structural relationships between domains. Another approach estimates the maximum likelihood that domains interact [22], [23]. Further ideas include overrepresented domain signatures [24], [25], domain combination [26], graph-theoretical methods [27], Bayesian networks [28], [29], support vector machines [30], and other probabilistic approaches [31], [32], [33].

Here, we introduce a novel method for the inference of protein interactions. Generalizing the complex relationships of interactions among proteins and their domain architectures, we conceptualize a maximum-specificity set cover procedure (MSSC), allowing us to determine sets of protein domain interactions which describe the presence of protein interactions to the largest, most specific extent. We utilize interaction networks of proteins in *S. cerevisiae* and their corresponding Pfam [34] domain architectures to determine probabilities of putative protein interactions, allowing for levels of sensitivity and specificity which at least match previous methods. As for quality, our predicted interactions correlate significantly with elevated levels of coexpression

• C. Huang, F. Morcos, D.Z. Chen, and J.A. Izaguirre are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556.

E-mail: {chuang1, amorcos, izaguirr}@nd.edu, chen@cse.nd.edu.

• S.P. Kanaan is with Accenture, 1710 N. Talman Ave., Chicago IL 60647. E-mail: simon.p.kanaan@accenture.com.

• S. Wuchty is with the Northwestern Institute of Complexity (NICO), Kellogg School of Business, Northwestern University, Evanston, IL 60208. E-mail: s-wuchty@northwestern.edu.

Manuscript received 20 Aug. 2004; revised 14 Nov. 2005; accepted 22 Feb. 2006; published online 31 Jan. 2007.

For information on obtaining reprints of this article, please send e-mail to: [tcbb@computer.org](mailto:tcbb@computer.org), and reference IEEECS Log Number TCBB-0095-0804. Digital Object Identifier no. XXX

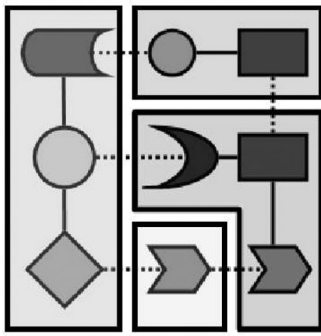


Fig. 1. The fundamental units of proteins (shaded areas) are the domains (geometrical figures), mediating a distinct structure and biological functionality. We assume that the underlying protein domain architectures facilitate the interactions among proteins, allowing us to design a novel method for the inference of protein interactions in *S. cerevisiae*.

as well as with low distances between GO terms of corresponding proteins. Focusing on biologically relevant examples, we show that MSSC reliably predicts previously unknown protein interactions in well studied molecular systems such as the 26S proteasome and RNA polymerase of *S. cerevisiae*.

## 2 MATERIALS AND METHODS

### 2.1 Protein Interactions

The first comprehensive, albeit weakly overlapping protein interaction maps of *S. cerevisiae* have been provided with the yeast-two-hybrid method [2], [3]. Currently, there exists a variety of yeast specific protein interaction databases. Many of them, such as MINT [35], MIPS [36], and BIND [37], collect experimentally determined protein interactions. PREDICTOME [38] and STRING [39] collect functional links between proteins, derived from genome scale two-hybrid sets, domain fusion events, phylogenetic history, and gene proximity. These databases lack an assessment of the data's quality. In contrast, the GRID database, a compilation of BIND, MIPS, and other data sets, as well as the DIP database [40], provides sets of manually curated protein-protein interactions in *S. cerevisiae*. Recently, Bader et al. introduced a novel method for the assessment of the quality of interactions utilizing other sources of information, including mRNA expression, genetic interactions, and database annotations. In particular, a logistic regression procedure allows the reliable validation of 47,783 experimentally obtained protein interactions for 4,627 proteins in yeast [5] by a confidence score which ranges from 0 to 1. We focus on 2,973 yeast proteins embedded in 11,368 interactions which score  $\geq 0.5$ . DIP is a subset of Bader et al.'s larger data set. With the confidence threshold used we have a data set of very similar size and quality compared to DIP.

### 2.2 Protein Complex Data

As a source of protein complex data, we utilized a set of minimally redundant, comprehensive complexes in yeast [41]. These data have been obtained by integrating large-scale experiments for the determination of yeast complexes [42], [43] using unsupervised clustering, allowing for 1,041 curated yeast protein clusters.

### 2.3 Protein Domains

For our analysis, we focused on domain data retrieved from the Pfam database, a reliable collection of multiple sequence alignments of protein families and profile hidden Markov models [34] (<http://pfam.wustl.edu>). We used Pfam version 10.0, which contains 6,190 fully annotated Pfam-A families. Pfam-B provides additional PRODOM-generated [44] alignments of sequence clusters in SWISSPROT and TrEMBL [45] that are not modeled in Pfam-A. In order to construct the Pfam domain architecture, we browsed swisspfam, a compilation of the domain structure of SWISSPROT and TrEMBL proteins according to Pfam.

#### 2.3.1 Microarray Data and Coexpression Correlation Coefficients

Genes with similar expression profiles are likely to encode interacting proteins [46], [47], [48]. Ideally, strongly coexpressed proteins should allow us to confirm the existence of predicted protein interactions. Yet, the noisy nature of experimental procedures for the determination of interactions often impedes this goal. However, strong correlations between gene expression data and interactions exist primarily among proteins which are related to permanent protein complexes. By downloading 1,051 expression profiles of yeast from the Stanford Microarray Database (SMD, <http://genome-www5.stanford.edu>), we calculated the Pearson's correlation coefficient  $r_P$  for each pair of interacting proteins. Provided that we find data for both proteins over  $m$  time points, the Pearson correlation coefficient is calculated by

$$r_P = \frac{\frac{1}{m} \sum_{i=1}^m x_i y_i - \bar{x} \bar{y}}{\sigma_i \sigma_j}, \quad (1)$$

where  $\bar{x}$  and  $\bar{y}$  are the sample means and  $\sigma_i$  and  $\sigma_j$  are the standard deviations of  $i$  and  $j$ .

#### 2.3.2 GO Annotation Data and GO Distance

For any two interacting proteins, we calculate an annotation-based distance between proteins, taking into account all Gene Ontology terms [49] (GO, <http://www.geneontology.org>) that are common to the pair and terms which are specific to each protein. Any two proteins can have several shared GO terms (common terms) and a variable number of terms specific for each protein (specific terms). This distance between interacting proteins  $i$  and  $j$  is based on the Czekanowski-Dice formula [50]:

$$d_{i,j} = \frac{|T_{GO}(i) \Delta T_{GO}(j)|}{|T_{GO}(i) \cup T_{GO}(j)| + |T_{GO}(i) \cap T_{GO}(j)|}. \quad (2)$$

In this formula,  $T_{GO}$  are the sets of the proteins' associated GO terms, while  $|T_{GO}|$  stands for their number of elements and  $\Delta$  is the symmetrical difference between two sets. This distance formula emphasizes the importance of the shared GO terms by giving more weight to similarities than to differences. Consequently, for two genes that do not share any GO terms, the distance value is 1, while, for two proteins sharing exactly the same set of GO terms, the distance value is 0.

## 2.4 Quality Measures

The quality of our interactions is assessed by comparing a set of predicted interactions  $P$ , which exceed a certain threshold of the prediction score, to a *testing set*  $T$ , a sample of known protein interactions. Formally, we define specificity as the ratio of the number of matched interactions between predictions  $P$  and the testing set,  $T$ , over the total number of predicted interactions in  $P$ ,  $S_p = \frac{|P \cap T|}{|P|}$ . In turn, we define sensitivity as the ratio of the number of matched interactions between  $P$  and  $T$  over the total number of observed interactions in the testing set  $T$ ,  $S_n = \frac{|P \cap T|}{|T|}$ . Thus, these metrics are dependent on the choice of the testing set as well as the prediction score threshold.

## 3 PREVIOUS PREDICTION METHODS

In order to have an estimate of the quality of our predictions, we compare the performances of our prediction method and previous methods. In the following, we will give a brief description of the most relevant algorithms that utilize protein interactions and their corresponding domain profiles to predict protein interactions in *S. cerevisiae*.

### 3.1 Association Method (AM)

The association method [24] assigns an interaction probability

$$\Pr(d_m, d_n) = \frac{I_{mn}}{N_{mn}}, \quad (3)$$

to each domain pair  $(d_m, d_n)$ .  $I_{mn}$  is the number of interacting protein pairs that contain  $(d_m, d_n)$ , and  $N_{mn}$  is the total number of protein pairs that contain  $(d_m, d_n)$ .

### 3.2 Maximum Likelihood Estimation (MLE)

The maximum likelihood estimation method (MLE) [22] assumes that two proteins interact if at least one pair of domains of the two proteins interacts. Assuming that interactions between different domain pairs are independent, the probability of a potential interaction between a protein pair  $(P_i, P_j)$  is

$$\Pr(P_{ij} = 1) = 1 - \prod_{(d_m, d_n) \in (P_i, P_j)} (1 - \lambda_{mn}), \quad (4)$$

where  $\lambda_{mn}$  denotes the probability that domain  $d_m$  interacts with domain  $d_n$ . Application of MLE [22] achieved 42.5 percent specificity and 77.6 percent sensitivity on a combined yeast protein interaction set compiled from [3], [2]. The Expectation-Maximization (EM) algorithm was used to maximize the likelihood. We implemented MLE and EM to test against MSSC. Details of the method and its implementation, and access to the source code are found in the supplementary material Web site, <http://ppi.cse.nd.edu>.

## 4 MAXIMUM SPECIFICITY SET COVER (MSSC)

Here, we present a novel method to predict protein-protein interactions. In our algorithm, we model protein interactions as being explained by the presence of distinct families of protein domain interactions. Assuming that interactions between different domain pairs are independent, we

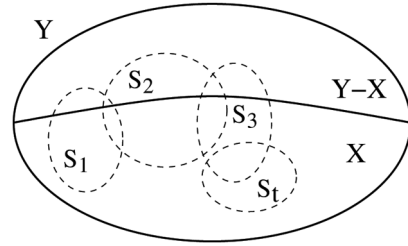


Fig. 2. The generalized set cover problem:  $X$  is a subset of  $Y$  and  $\mathcal{F} = \{S_i, 1 \leq i \leq t\}$  is a family of subsets of  $Y$ .

conceptualize the intricate interplay between protein interactions, architecture of domains, and their interactions as a generalized weighted set-cover problem, aiming to find a set of domain pairs which “covers” the given protein-protein interactions to the largest extent.

### 4.1 General Set Cover Problem

Suppose  $X$  is a finite set and  $\mathcal{F}$  is a family of subsets of  $X$  that can cover  $X$ , i.e.,  $X = \bigcup_{S \in \mathcal{F}} S$ . The set-cover problem is to find a subset  $\mathcal{C}$  of  $\mathcal{F}$  to cover  $X$ ,

$$X = \bigcup_{S \in \mathcal{C}} S, \quad (5)$$

and  $\mathcal{C}$  is also required to satisfy certain conditions according to specific problems. For example, the minimum exact set-cover (MESCC) problem requires that  $\sum_{S \in \mathcal{C}} |S|$  is minimized while the minimum set-cover (MSC) problem is to find a  $\mathcal{C}$  with minimum cardinality  $|\mathcal{C}|$  [51], [52].

We generalize the set-cover problem by enclosing  $X$  into a bigger set  $Y$  (Fig. 2). Suppose  $Y$  is a finite set,  $X \subseteq Y$  and  $\mathcal{F}$  is a family of subsets of  $Y$  that can cover  $X$ , i.e.,  $X \subseteq \bigcup_{S \in \mathcal{F}} S$ . The generalized set-cover problem is to find a subset  $\mathcal{C}$  of  $\mathcal{F}$  to cover  $X$ ,

$$X \subseteq \bigcup_{S \in \mathcal{C}} S, \quad (6)$$

and  $\mathcal{C}$  is also required to satisfy certain conditions according to different specific problems, as before.

### 4.2 Concept of Protein Interactions as a Set Cover Problem

Solving the protein-protein interaction problem, we map the complex relationships between interactions of proteins and their domain architectures to a set-cover problem. The experimentally known protein-protein interaction network is modeled by a graph  $G = (P, E)$ , where  $P$  is the set of proteins (vertices), while  $E$  is the set of interactions (edges). Formally, the protein interaction network is represented as a set-cover problem by defining

$$Y = \{\text{all protein pairs } (P_i, P_j) \mid P_i, P_j \in P\},$$

$$X = \{\text{protein pairs } (P_i, P_j) \mid P_i \text{ interacts with } P_j \text{ in } G\},$$

and  $\mathcal{F}$  is the set of all domain pairs  $(d_m, d_n)$  (see Fig. 3 for a schematic representation of these relations). A domain pair  $(d_m, d_n)$  is viewed as a subset of  $Y$ . Specifically, if a protein pair  $(P_i, P_j)$  (an element in  $X$ ) contains  $(d_m, d_n)$ , then  $(P_i, P_j)$  belongs to the subset  $(d_m, d_n)$ . In Fig. 4, we show an illustration of all possible domain pairs that can cover a

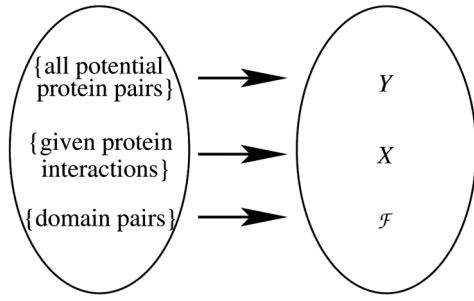


Fig. 3. Transforming protein interactions into a set cover problem: The big set  $Y$  is taken to be the set of all potential protein pairs, the subset  $X$  is taken to be the set of the given protein-protein interactions, and the family  $\mathcal{F}$  is taken to be the set of all the domain pairs.

protein interaction, elements in a domain pair. Some of the elements (protein pairs) are given interactions. Suppose we find a subset  $\mathcal{C}$  of  $\mathcal{F}$  to cover every element  $(P_i, P_j)$  in  $X$ . An element in  $\mathcal{C}$  corresponds to a domain pair  $(d_m, d_n)$ . If  $(d_m, d_n)$  covers  $(P_i, P_j)$ , then the two proteins  $P_i$  and  $P_j$  contain  $d_m$  and  $d_n$ , respectively; so,  $(d_m, d_n)$  can be used to represent the interaction between  $P_i$  and  $P_j$ . Therefore, we also have a set of domain pairs to represent the protein network  $G$ . On the other hand, suppose there is a set  $D$  of domain pairs to represent the network  $G$ . For every element  $(P_i, P_j)$  in  $X$ , there is a domain pair  $(d_m, d_n)$  from  $D$  to represent the interaction between  $P_i$  and  $P_j$ . Since  $(d_m, d_n)$  can be viewed as an element in  $\mathcal{F}$ , the collection  $\mathcal{C}$  of all the domain pairs from  $D$  is a subset of  $\mathcal{F}$ , and  $\mathcal{C}$  covers  $X$ .

### 4.3 MSSC Approach

There are many ways to choose domain pairs to represent the protein interaction network. AM simply uses all possible domain pairs to explain protein-protein interactions, i.e., it uses  $\mathcal{F}$  to cover  $X$ , resulting in a very low specificity [22]. We are interested in finding a subset of domain pairs which allows us to represent the protein-protein interaction network by maximizing both the specificity and sensitivity on the training set.

The MSSC problem is to find a subset  $\mathcal{C}$  of  $\mathcal{F}$  to cover  $X$  such that

$$m(\mathcal{C}) := \sum_{S \in \mathcal{C}} |S - X| \quad (7)$$

is minimized.

We can see that MSSC allows the subcover  $\mathcal{C}$  to cover the overlap with  $X$ , while the overlap with  $Y - X$  (outside  $X$ ) is

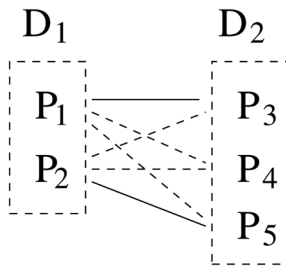


Fig. 4. Each domain in the domain pair  $(D_1, D_2)$  is contained in a list of proteins.  $(D_1, D_2)$  is a subset of  $Y$ , the set of all protein pairs. As such,  $(D_1, D_2)$  covers those protein pairs, where some indeed interact (solid lines). Our algorithm seeks the best cover that explains the interactions to the largest extent.

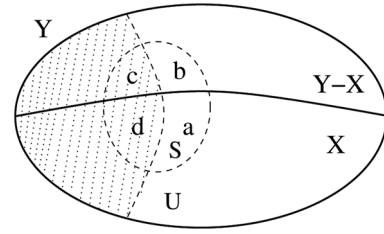


Fig. 5. The shaded area is already covered by  $C$ .  $U$  is the unshaded area in  $X$ . The candidate set  $S$  is divided into four parts, a, b, c and d. MSSC chooses a set  $S$  with the minimum  $\frac{b+c}{a}$ . The greedy algorithm for MSSC allows overlapping of subcover inside  $X$ , actually increasing the interaction probability for a protein pair.

minimized, an optimization constraint that allows MSSC to maximize the specificity of the chosen cover because false positives are considered to appear only in  $Y - X$ . Allowing a fast and efficient computation of the required set, we choose a greedy algorithm (Algorithm 2) as the basic routine for MSSC. We compared MSSC to MSC and MESC and we found that MSSC gives vastly superior results to MSC and similar or better results than MESC. In this paper, we only show results of MSSC.

**Algorithm 1: Greedy algorithm for MSSC.** Allows the determination of a subset of domain pairs ( $\mathcal{F}$ ) that covers the underlying training set of protein interactions ( $X$ ) with maximum specificity since false positives are considered to appear in  $Y - X$ , where  $Y$  is the set of all protein pairs.  $U$  represents the uncovered part of  $X$ .  $\mathcal{E}$  is the subset of  $\mathcal{F}$  that has not been chosen by the algorithm.

GREEDY\_MSSC( $Y, X, \mathcal{F}$ )

$U \leftarrow X$

$\mathcal{E} \leftarrow \mathcal{F}$

$\mathcal{C} \leftarrow \emptyset$

**while**  $U \neq \emptyset$

**do** select an  $S \in \mathcal{E}$  with the minimum  $\frac{|S-X|}{|S \cap U|}$

(a tie is broken by  $|S \cap U|$ )

$U \leftarrow U - S$

$\mathcal{E} \leftarrow \mathcal{E} - \{S\}$

$\mathcal{C} \leftarrow \mathcal{C} \cup \{S\}$

**return**  $\mathcal{C}$

In this algorithm, at each step when a subset needs to be chosen, we choose the one whose ratio between the part outside  $X$  and the part inside  $U$  is minimized (Fig. 5).

The above greedy algorithm is just an approximation and the solution found by it has the following relationship with the optimal solution of MSSC:

**Theorem 4.1.** Suppose  $\mathcal{C}_a$  is the approximation of MSSC found by the above greedy algorithm and  $\mathcal{C}_o$  is an optimal subcover for MSSC. Let  $k = \max_{S \in \mathcal{F}} |S|$ . If  $m(\mathcal{C}_o) = 0$ , then  $m(\mathcal{C}_a) = 0$ ; otherwise, we have

$$\frac{m(\mathcal{C}_a)}{m(\mathcal{C}_o)} \leq [\ln(k-1) + 1]. \quad (8)$$

TABLE 1  
Data Sets Used to Generate Figures in This Paper

Number	Data set	# Proteins	# Interactions	Remarks
I	Bader <i>et al.</i>	2,973	11,368	Confidence $\geq 0.5$ ; Used for Figs. 6, 8b, and 9
II	Protein complexes	1,782	6,680	1,041 curated complexes; Used for Figs. 7 and 8a
III	Physical interactions	1,851	2,353	Used for Figs. 7 and 8a
IV	26S Proteasome	36	279	Used for Fig. 9
V	RNA Polymerase II	32	226	Used for Fig. 9

The first three rows are partitioned into disjoint training and testing sets using the ratio 80 percent:20 percent, respectively. The last two entries are testing sets only, which use as a training set the whole of the first row.

The proof of Theorem 4.1 can be found on the supplementary material Web site and in [53]. The theorem shows the relationship between the approximation by *GREEDY\_MSSC* and an optimal solution. If  $k$  is small, the difference between them is small too. In this theorem,  $k$  is the maximum number of elements a subset can have and it corresponds to the maximum number of protein pairs that contain a domain pair in the protein network.

When  $X = Y$ , MSSC is reduced to MSC, which is well known to be NP-hard. In the case of MSC, a logarithmic approximation is the best known approximation.

#### 4.4 Prediction

Once the domain pairs are chosen by MSSC, each pair is assigned the same interaction probability as in AM (3), while unchosen domain pairs are given an interaction probability 0. Subsequently, (4) is used to calculate the interaction score for each putative protein pair (between 0 and 1).

## 5 RESULTS

In Table 1, we briefly describe the input sets used for training and testing sets that are used to generate the

figures in this section. These data sets and the predictions are available from <http://ppi.cse.nd.edu>. We refer to these data sets in the following sections by the number in this table.

### 5.1 Comparison of Performance

We compare the ability of MSSC to predict protein-protein interactions against AM and MLE utilizing Data Set I (see also Section 2). In particular, we pool 80 percent of these protein interactions in a training set, while the remainder serves as the test set. We use the training set of protein interactions to determine a set of potential domain interactions which allow the description of the interactions in the testing set to the largest specific extent (see Section 2). Fig. 6a shows that the predictions using MSSC and MLE are very similar, highlighting that, in regions of elevated specificity, MSSC and MLE predictions provide higher sensitivity than AM. We also observe that MLE does not have any predicted set when the specificity is high enough (between 60 percent and 100 percent). However, the main advantage of MSSC is speed. MSSC only needs 73 seconds to provide these results. In contrast, MLE ran for more than

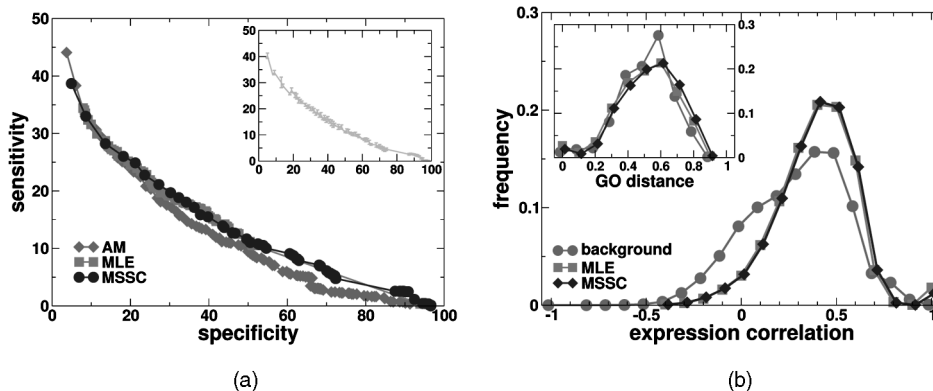


Fig. 6. (a) Utilizing 80 percent of high quality pairwise interactions of yeast as the training set while the remainder provides the testing set, MLE and MSSC show similar specificity versus sensitivity curves. Randomly sampling 80 percent out of the pool of pairwise interactions (Data Set I in Table 1) as training sets and averaging over 10 realizations, we observe that the fluctuations in the predictions utilizing MSSC are small, and the differences between MSSC and MLE, on one hand, and AM, on the other hand, are statistically significant (inset). (b) Evaluating the quality of predictions with the MSSC and MLE algorithms, we utilize 80 percent of yeast interactions of Bader *et al.* (Data Set I) as a training set while the remainder serves as testing set. Compared to a set of random chosen pairs of yeast proteins, we observe that the distributions of coexpression correlation coefficients allowing the characterization of the quality of predicted interactions significantly shift toward high levels of coexpression ( $t_{MLE} = 30.6, P \ll 10^{-15}$ ;  $t_{MSSC} = 29.7, P \ll 10^{-15}$ ). Despite significant different means ( $t = 4.4, P < 10^{-7}$ ), the differences between predictions as of the MSSC and MLE algorithm only differ marginally. Similarly, the comparison of GO distances to a random background distribution differs significantly (inset,  $t_{MLE} = 28.1, P \ll 10^{-15}$ ;  $t_{MSSC} = 45.9, P \ll 10^{-15}$ ) while we find no large differences in the distributions of GO distances of predictions between MSSC and MLE algorithm ( $t = 7.9, P < 10^{-14}$ ).

6 hours; this is due to the iterative nature of the MLE method.

In the inset of Fig. 6a, we demonstrate that sensitivity and specificity obtained from randomly sampling training and testing sets are widely invariant. In particular, we randomly sample Data Set I interactions into two disjoint sets (80 percent:20 percent). Taking the larger sample as the training set and averaging over 10 realizations, we obtain small fluctuations in the corresponding specificity/sensitivity behavior of the predictions (inset, Fig. 6a). This shows that the difference between AM and MSSC or MLE is statistically significant.

The low sensitivity in the case of disjoint training and testing sets is rooted in the fact that interactions in the training and testing set do not share many domain pairs. Although AM allows many more domain interactions by choosing all the domain pairs between the interactions in the training set, this algorithm only reaches 42 percent specificity at best. In other words, about 58 percent of interactions in the testing set do not have any domain pairs in common with the training set, encouraging Han et al. [26] to exclude those to get higher sensitivity. A different assessment of the prediction quality is the tendency of interactions toward coexpression and functional homogeneity of the interacting proteins. As a background set, we randomly sampled 100,000 protein pairs. Utilizing coexpression data of yeast proteins (see Section 2), we calculated coexpression correlation coefficients of each pair, allowing us to obtain a bell-shaped frequency distribution peaking around 0.5. If there exists a correlation between the presence of an interaction between a pair of proteins and their coexpression, we expect a shift to higher expression coefficients. In Fig. 6b, both MSSC and MLE significantly provide an enrichment of coexpressed interacting proteins. Assuming that the observed distributions have roughly the same variance, we apply a Student's t-test to uncover possibly different means of the predicted coexpression profiles. Compared to the random background distribution, we find that both sets of predictions significantly differ in their means, while the differences between predictions, as of the MSSC and MLE algorithm only differ marginally. Similarly, interacting proteins tend to share Gene Ontology (GO) terms, another observation which qualifies as a measure of the predictions quality. Indeed, t-student test scores indicate that the comparison of GO distances (see Section 2) of the predicted interactions differ significantly from the random sample of protein pairs, while we find no large differences in the distributions of GO distances of predictions between MSSC and MLE (inset, Fig. 6b).

## 5.2 Results with Complex or Physical Interactions

We assume that the nature of the underlying interactions in our training set might have a considerable impact on the predictive ability of our algorithm. In particular, we expect that interactions which only occur in protein complexes will allow us to obtain different levels of quality compared to predictions obtained with physical interactions. We use Data Set II, a well-integrated and curated set of 1,041 protein complexes that have been compiled from large-scale experiments [42], [43]. Data Set II is obtained by splitting Data Set I into a set of interactions that exclusively occur in

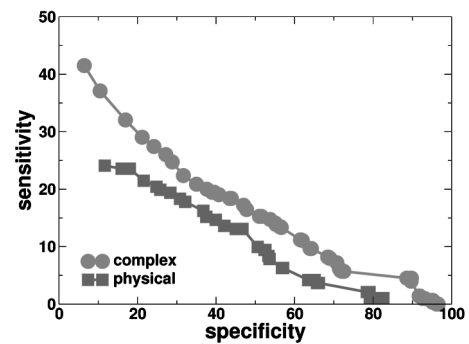


Fig. 7. Differentiating between protein interactions that belong to protein complexes and physical protein interactions, we utilize 80 percent of each set as the training set while we test our predictions with the remaining interactions in each set. Interestingly, we observe that predictions obtained with complex interactions outscore their physical counterpart in terms of sensitivity and specificity.

complexes only. We consider the remainder as physical interactions (Data Set III). Utilizing samples of 80 percent of interactions in each subset as training data while the remainder serve as testing sets, we observe that protein interactions we obtained from complexes provide higher sensitivity than physical interactions (Fig. 7). In Fig. 8a, we compare the predicted interactions sets to the random chosen pairs of yeast proteins: Both distributions of coexpression correlation coefficients significantly shift toward high levels of coexpression. In particular, predictions obtained with physical interactions slightly shift stronger toward higher coexpression. As for differences in GO annotations, predictions obtained with complex interactions differ more from a random background distribution than predictions obtained with physical interactions. Predictions also provide considerable differences between each other, allowing us to conclude that the nature of the underlying training set of interactions impacts the functional homogeneity of predictions.

Focusing on predicted interactions that score above a confidence score greater than 0.5 (see Section 2.1), we observe a much improved picture (Fig. 8b). In particular, we observe strong shifts toward higher coexpression, results which are supported by significant t-test values when compared to the random background distribution. Comparing these predictions with each other, we observe that the means of the distributions are similar. We obtain a similar picture with respect to GO distances between the predicted interacting proteins, where both sets show a strong shift toward low GO distances (inset, Fig. 8b).

## 5.3 26S Proteasome and RNA Polymerase II

As concrete examples that offer support for the strength of our predictions, we chose 26S proteasome and RNA polymerase II holoenzyme as representative examples (Data Sets IV and V, respectively). Both protein complexes are well-studied molecular systems for which the three-dimensional structure is well known [48]. In particular, we chose the complete Data Set I as training set, allowing us to find interactions that are expected (being in the training set) and predicted. In Fig. 9, we show interaction maps of the respective examples. The 26S proteasome consists of the 20S

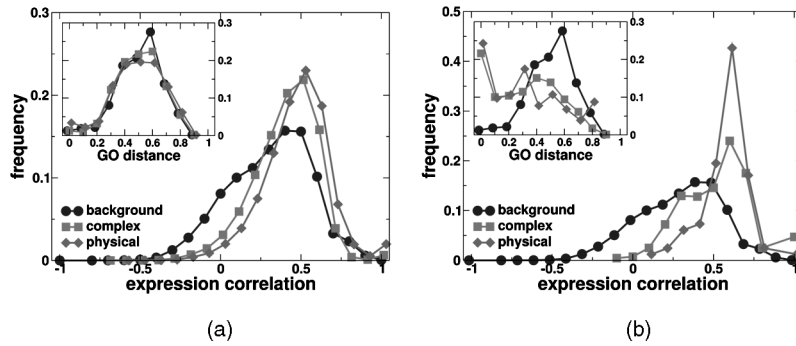


Fig. 8. We expect that the nature of the underlying protein interactions in yeast will have a significant effect on the quality of the predictions with the MSSC algorithm. In particular, we chose 80 percent of physical and complex interactions as training sets while the remainder served as testing sets. (a) We consider all predictions regardless of the corresponding interaction probability. Compared to a set of randomly chosen pairs of yeast proteins, we observe that both distributions of coexpression correlation coefficients significantly shift toward high levels of coexpression ( $t_{physical} = 22.9, P \ll 10^{-15}; t_{complex} = 50.7, P \ll 10^{-15}$ ). Comparing predictions obtained with physical and complex interactions, we observe that predictions obtained with physical interactions exhibit a slightly stronger shift toward higher coexpression ( $t = 13.4, P \ll 10^{-15}$ ). As for differences in GO annotations, we observe that predictions obtained with complex interactions differ more from a random distribution than predictions obtained with physical interactions ( $t_{physical} = 6.7, P < 10^{-11}; t_{complex} = 33.3, P \ll 10^{-15}$ ). Predictions also provide considerable differences between each other ( $t = 12.7, P \ll 10^{-15}$ ). (b) Accounting for predicted interactions that score above  $P \geq 0.5$ , we observe strong shifts toward higher coexpression, results which are supported by significant t-test values when compared to the random background distribution ( $t_{physical} = 9.0, P \ll 10^{-15}; t_{complex} = 20.6, P \ll 10^{-15}$ ). Comparing these predictions with each other, we observe that the means of the distributions are similar ( $t = 1.9, P = 5.6 \times 10^{-2}$ ). We obtain a similar picture with respect to GO distances between the predicted interacting proteins, where both sets show a strong shift toward low GO distances (inset). While we find significant differences to the random background distribution ( $t_{physical} = 5.5, P = 3.4 \times 10^{-8}; t_{complex} = 7.2, P = 6.3 \times 10^{-13}$ ), distributions of predicted interactions are very similar ( $t = 0.15, P = 0.87$ ), allowing us to conclude that the nature of the underlying training set of interactions impacts the functional homogeneity of predictions.

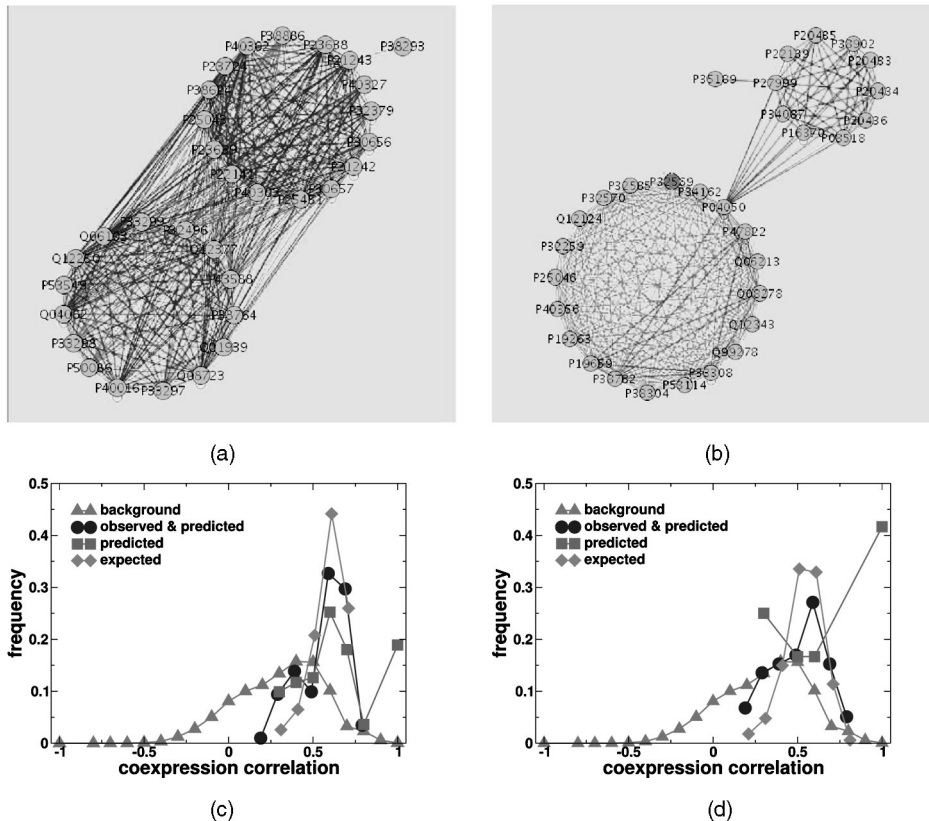


Fig. 9. Predicted and expected interactions in the (a) 26S proteasome and (b) RNA polymerase II. In particular, we chose the complete set of high quality interactions as a training set, allowing us to find interactions that are both expected (i.e., in the training set) and predicted (green lines), expected but not predicted (yellow), and predicted but not expected (red lines). In particular, interactions in both subunits of the 26S proteasome, in the 20S and 19/22S proteasome, and among the subunits of the RNA polymerase II are well predicted by MSSC (202 matches, 77 missed, and 101 new). In contrast, we find many predicted interactions between the subunits of the proteasome while we do not find many interactions among the larger Kornberg’s mediator (SRB) complex of the RNA polymerase II holoenzyme (59 matches, 167 missed, and 12 new). In the lower panel, we show coexpression correlations of the different types of interactions, illustrating the strong shift to coexpression for these complexes compared to a background distribution, in agreement with results from [48]. Interaction graphs in panels (a) and (b) were created with *Cytoscape*.

and 19/22S subunit, a distinction that is well reflected by an accumulation of protein interactions that are both expected and predicted by MSSC. MSSC also predicts many interactions between the two subunits. As for coexpression correlations of the expected and predicted interactions, we find that the coexpression patterns not only resemble each other but assemble at high values. In the case of the RNA polymerase II complex, we find excellent agreement of our predictions with interactions that appear between the subunits of the polymerase. However, we do not find many interactions among the larger Kornberg's mediator (SRB) complex, a consequence of the absence of any information about the Pfam domain composition of a large majority of involved proteins. Coexpression patterns of the different types of links resemble each other. Links that have been predicted by MSSC but are not expected peak at high values of coexpression, which signals the presence of protein self-interactions.

## 6 DISSEMINATION

We created a Web portal for protein-protein interaction, <http://ppi.cse.nd.edu>. The portal provides a means for predicting and analyzing interactions using the prediction methods MSSC, AM, and MLE. All input and output data sets described in this paper and the source code can be downloaded too. The portal allows the user to create or download protein interaction data sets, calculate quality metrics for predicted protein pairs, predict interactions, and analyze the PPI program output. This consists of the ability to view and download a Sensitivity versus Specificity plot, and search individual protein names to get a list of all of its interactors.

## 7 DISCUSSION

In this paper, we showed a new way of integrating protein interaction and domain data, ultimately allowing us to predict previously unknown protein interactions. By design, our MSSC approach selects a set of domain pairs that both cover the experimental observations and maximize the specificity in the training set. Our results indicate that there is a strong correlation between high specificity in high quality training sets and high specificity in realistic testing sets.

In comparison, MSSC at least reaches the level of specificity and sensitivity of MLE and it is faster than MLE. The specificity and sensitivity of MSSC (and MLE) are consistently higher than those of AM. Even though improved versions of AM exist, we think that MLE is, in a way, the optimal algorithm, given our assumptions. MSSC is an attractive alternative, of similar quality but faster execution time.

As a proof of concept, we observe that predicted interactions tend to be coexpressed and exhibit small distances between the involved proteins GO terms. Although our results are encouraging, we are aware of limitations. As the example of the 26S proteasome shows, we observe that our algorithm only allows predictions between proteins with a well-known domain architecture as well as known interactions among the respective domains. In our approach, we do not account for any three-dimensional information but

infer potential domain interactions by counting the occurrence of all possible domain pairs that the domain architectures of interacting proteins imply, a method that, by definition, risks an elevated level of noise in the determination of potential domain interactions. Accordingly, the error proneness of protein interactions in the respective training sets is another source of potential noise, impacting the quality of predictions. As such, a further improvement of the algorithm will focus on the systematic integration of high quality interaction data as well as selection steps of reliable domain interactions in order to ensure highly reliable predictions. Once large-scale protein interaction sets of organisms other than *S. cerevisiae* are available, we expect that our algorithm will significantly contribute to the elucidation of complete organism-specific interactomes.

## ACKNOWLEDGMENTS

J.A. Izaguirre, C. Huang, and S.P. Kanaan were partially funded by US National Science Foundation (NSF) grants IBN-0083653, IBN-0313730, and ACI-0135195. F. Morcos was partially funded by a Kellogg fellowship. D.Z. Chen was supported in part by US NSF Grant CCF-0515203. The simulations were run in a cluster funded by Notre Dame's high performance cluster grant to J.A. Izaguirre. The authors give special thanks to Mr. Lance Gallop and Mr. Kyle Marks for developing the ppi Web portal.

## REFERENCES

- [1] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A Comprehensive Two-Hybrid Analysis to Explore the Yeast Protein Interactome," *Proc. Nat'l Academy of Science USA*, vol. 98, no. 8, pp. 4569-4574, 2001.
- [2] T. Ito, K. Tashiro, S. Muta, R. Ozawa, T. Chibba, M. Nishizawa, K. Yamamoto, S. Kuhara, and Y. Sakaki, "Towards a Protein-Protein Interaction Map of the Budding Yeast: A Comprehensive System to Examine Two-Hybrid Interactions in All Possible Combinations between the Yeast Proteins," *Proc. Nat'l Academy of Science USA*, vol. 97, no. 3, pp. 1143-1147, 2000.
- [3] B. Schwikowski, P. Uetz, and S. Fields, "A Network of Protein-Protein Interactions in Yeast," *Nature Biotechnology*, vol. 18, pp. 1257-1261, 2000.
- [4] P. Uetz, L. Giot, G. Cagney, T. Mansfield, R. Judson, J. Knight, D. Lockshorn, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamar, M. Yang, M. Johnston, S. Fields, and J. Rothberg, "A Comprehensive Analysis of Protein-Protein Interactions of *Saccharomyces cerevisiae*," *Nature*, vol. 403, pp. 623-627, 2000.
- [5] J.S. Bader, D. Chaudhuri, and J. Chant, "Gaining Confidence in High-Throughput Protein Interaction Networks," *Nature Biotechnology*, vol. 22, pp. 78-85, 2004.
- [6] P. Bork, L. Jensen, C. von Mering, A. Ramani, and E. Marcotte, "Protein Interaction Networks from Yeast to Human," *Current Opinion on Structural Biology*, vol. 14, pp. 292-299, 2004.
- [7] M. Vidal, "Interactome Modelling," *FEBS Letters*, vol. 579, pp. 1834-1838, 2005.
- [8] D. LaCount, M. Vignali, R. Chettier, A. Phansalkar, R. Bell, J. Hesselberth, L. Schoenfeld, S.S.I. Ota, C. Kurschner, S. Fields, and R. Hughes, "A Protein Interaction Network of the Malaria Parasite *Plasmodium falciparum*," *Nature*, vol. 438, pp. 103-107, 2005.
- [9] A. Walhout, R. Sordella, X. Lu, J. Hartley, G. Temple, M. Brasch, N. Thierry-Mieg, and M. Vidal, "Protein Interaction Mapping in *C. elegans* Using Proteins Involved in Vulval Development," *Science*, vol. 287, pp. 116-122, 2000.
- [10] S. Li, C. Armstrong, N. Bertin, H. Ge, S. Milstein, M. Boxem, P.-O. Vidalain, J.-D. Han, A. Chesneau, and T. Ha, "A Map of the Interactome Network of the Metazoan *C. Elegans*," *Science*, vol. 303, pp. 540-543, 2004.

- [11] L. Giot, J. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. Hao, C. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carroll, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. Stanyon, R. Finley Jr., K. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. Shinkets, M. McKenna, J. Chant, and J. Rothberg, "A Protein Interaction Map of *Drosophila melanogaster*," *Science*, vol. 302, pp. 1727-1736, 2004.
- [12] J.-F. Rual et al., "Towards a Proteome-Scale Map of the Human Protein-Protein Interaction Network," *Nature*, vol. 437, pp. 1173-1178, 2005.
- [13] A. Enright, I. Iliopoulos, N. Kyripides, and C. Ouzounis, "Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events," *Nature*, vol. 402, pp. 86-90, 1999.
- [14] E. Marcotte, M. Pellegrini, M. Thompson, T. Yeates, and D. Eisenberg, "A Combined Algorithm for Genomewide Prediction of Protein Function," *Nature*, vol. 402, pp. 83-86, 1999.
- [15] M. Pellegrini, E. Marcotte, M. Thompson, D. Eisenberg, and T. Yeates, "Assigning Protein Functions by Comparative Genome Analysis: Protein Phylogenetic Profiles," *Proc. Nat'l Academy of Sciences USA*, vol. 96, pp. 4285-4288, 1999.
- [16] E. Marcotte, M. Pellegrini, H.-L. Ng, D. Rice, T. Yeates, and D. Eisenberg, "Detecting Protein Function and Protein-Protein Interactions from Genome Sequences," *Science*, vol. 285, pp. 751-753, 1999.
- [17] D. Ekman, J.F. -S. A. K. Björklund, and E. Elofsson, "Multi-Domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions," *J. Molecular Biology*, vol. 348, pp. 231-243, 2005.
- [18] J. Wojcik and V. Schächter, "Protein-Protein Interaction Map Inference Using Interacting Domain Profile Pairs," *Bioinformatics*, vol. 17, pp. 2965-3055, 2001.
- [19] J. Espadaler, R.J.O. Romero-Isart, and B. Oliva, "Prediction of Protein-Protein Interactions Using Distant Conservation of Sequence Patterns and Structure Relationships," *Bioinformatics*, vol. 21, pp. 3360-3368, 2005.
- [20] P. Aloy, B. Böttcher, H. Ceulemans, C. Leutwein, C. Mellwig, S. Fischer, A.-C. Gavin, P. Bork, G. Superti-Furga, L. Serrano, and R. Russell, "Structure-Based Assembly of Protein Complexes in Yeast," *Science*, vol. 303, pp. 2026-2029, 2004.
- [21] A. Stein, R. Russell, and P. Aloy, "3did: Interacting Protein Domains of Known Three-Dimensional Structure," *Nucleic Acids Research*, vol. 33, pp. D413-D417, 2005.
- [22] M. Deng, S. Mehta, F. Sun, and T. Cheng, "Inferring Domain-Domain Interactions from Protein-Protein Interactions," *Genome Research*, vol. 12, pp. 1540-1548, 2002.
- [23] I. Iossifov, M. Krauthammer, C. Friedman, V. Hatzivassiloglou, J. Bader, K. White, and A. Rzhetsky, "Probabilistic Inference of Molecular Networks from Noisy Data Sources," *Bioinformatics*, vol. 20, pp. 1205-1213, 2004.
- [24] E. Sprinzak and H. Margalit, "Correlated Sequence-Signature as Markers of Protein-Protein Interaction," *J. Molecular Biology*, vol. 311, pp. 681-692, 2001.
- [25] S.M. Gomez, W.S. Noble, and A. Rzhetsky, "Learning to Predict Protein-Protein Interactions from Protein Sequences," *Bioinformatics*, vol. 19, pp. 1875-1881, 2003.
- [26] D. Han, H.-S. Kim, J. Seo, and W. Jang, "A Domain Combination Based Probabilistic Framework for Protein-Protein Interaction Prediction," *Genome Informatics*, vol. 14, pp. 250-259, 2003.
- [27] D. Goldberg and F. Roth, "Assessing Experimentally Derived Interactions in a Small-World," *Proc. Nat'l Academy of Sciences USA*, vol. 100, no. 8, pp. 4372-4376, 2003.
- [28] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. Krogan, S. Chung, A. Emili, M. Snyder, J. Greenblatt, and M. Gerstein, "A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data," *Science*, vol. 302, pp. 449-453, 2003.
- [29] N. Nariyai, S. Kim, S. Imoto, and S. Miyano, "Using Protein-Protein Interactions for Refining Gene Networks Estimated from Microarray Data by Bayesian Networks," *Proc. Pacific Symp. Biocomputing*, pp. 336-347, 2004.
- [30] I. Albert and R. Albert, "Conserved Network Motifs Allow Protein-Protein Interaction Prediction," *Bioinformatics*, vol. 20, pp. 3346-3352, 2004.
- [31] S. Gomez, S. Lo, and A. Rzhetsky, "Probabilistic Prediction of Unknown Metabolic and Signal Transduction Networks," *Genetics*, vol. 159, pp. 1291-1298, 2001.
- [32] A. Tong et al., "A Combined Experimental and Computational Strategy to Define Protein Interaction Networks for Peptide Recognition Modules," *Science*, vol. 295, pp. 321-324, 2002.
- [33] D. Rhodes, S. Tomlins, S. Varambally, V. Mahavisno, T. Barrette, S. Kalyana-Sundaram, D. Ghosh, A. Pandey, and A. Chinnaiyan, "Probabilistic Model of the Human Protein-Protein Interaction Network," *Nature Biotechnology*, vol. 23, pp. 951-959, 2005.
- [34] A. Bateman, L. Coin, R. Durbin, R. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. Sonnhammer, D. Studholme, C. Yeats, and S. Eddy, "The Pfam Protein Families Database," *Nucleic Acids Research*, vol. 32, pp. D138-D141, 2004.
- [35] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "MINT—A Molecular INteraction Database," *FEBS Letters*, vol. 513, pp. 135-140, 2002.
- [36] H.W. Mewes, U.B. D. Frishman, G. Mannhaupt, K. Mayer, M. Mokrejs, M.M. B. Morgenstern, S. Rudd, and B. Weil, "MIPS: A Database for Genomes and Protein Sequences," *Nucleic Acids Research*, vol. 30, pp. 31-34, 2002.
- [37] G. Bader, I. Donaldson, C. Wolting, B. Ouellette, T. Pawson, and C. Hogue, "BIND—the Biomolecular Interaction Network Database," *Nucleic Acids Research*, vol. 29, pp. 242-245, 2001.
- [38] J. Mellor, I. Yanai, K. Clodfelter, J. Mintseris, and C. DeLisi, "Predictome: a Database of Putative Functional Links between Proteins," *Nucleic Acids Research*, vol. 30, pp. 306-309, 2002.
- [39] C. vonMering, M. Huynen, D. Jaeggi, P.B.S. Schmidt, and B. Snel, "STRING: a Database of Predicted Functional Associations Between Proteins," *Nucleic Acids Research*, vol. 31, pp. 258-261, 2003.
- [40] I. Xenarios, L. Salwinski, X. Duan, P. Higney, S.-M. Kim, and D. Eisenberg, "DIP, the Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions," *Nucleic Acids Research*, vol. 30, pp. 303-305, 2002.
- [41] R. Krause, C. vonMering, and P. Bork, "A Comprehensive Set of Protein Complexes in Yeast: Mining Large Scale Protein-Protein Interaction Screens," *Bioinformatics*, vol. 19, pp. 1901-1908, 2003.
- [42] Y. Ho et al., "Systematic Identification of Protein Complexes in *Saccharomyces cerevisiae* by Mass Spectrometry," *Nature*, vol. 415, pp. 180-183, 2002.
- [43] A. Gavin, M. Bösch, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A.-M. Michon, C.-M. Cruciat, M. Remor, C. Böfer, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M.-A. Heurtier, R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes," *Nature*, vol. 415, pp. 141-147, 2002.
- [44] F. Corpet, F. Servant, J. Gouzy, and D. Kahn, "ProDom and ProDom-CG: Tools for Protein Domain Analysis and Whole Genome Comparisons," *Nucleic Acids Research*, vol. 28, no. 1, pp. 267-269, 2000.
- [45] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT Protein Knowledgebase and Its Supplement TrEMBL in 2003," *Nucleic Acids Research*, vol. 31, pp. 365-390, 2003.
- [46] A. Grigoriev, "A Relationship between Gene Expression and Protein Interactions on the Proteome Scale: Analysis of the Bacteriophage T7 and the Yeast *Saccharomyces cerevisiae*," *Nucleic Acids Research*, vol. 29, pp. 3513-3519, 2001.
- [47] H. Ge, L. Ziu, G. Church, and M. Vidal, "Correlation between Transcriptome and Interactome Mapping Data from *Saccharomyces cerevisiae*," *Nature Genetics*, vol. 29, pp. 482-486, 2001.
- [48] R. Jansen, D. Greenbaum, and M. Gerstein, "Relating Whole-Genome Expression Data with Protein-Protein Interactions," *Genome Research*, vol. 12, pp. 37-42, 2002.
- [49] "The Gene Ontology (GO) Database and Information Resource," *Nucleic Acids Research*, vol. 32, pp. D258-D261, G.O. Consortium, 2004.
- [50] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq, "GOToolBox: Functional Analysis of Gene Data Sets Based on Gene Ontology," *Genome Biology*, vol. 5, no. 12, pp. 1901-1908, 2004.

- [51] D.S. Johnson, "Approximation Algorithms for Combinatorial Problems," *J. Computer System Science*, vol. 9, pp. 256-278, 1974.
- [52] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms*, second ed. McGraw-Hill, 2001.
- [53] C. Huang, "Multiscale Computational Methods for Morphogenesis and Algorithms for Protein-Protein Interaction Inference," PhD dissertation, Dept. of Computer Science and Eng., Univ. of Notre Dame, July 2005, <http://etd.nd.edu/ETD-db/theses/available/etd-07212005-085435/>.



**Chengbang Huang** received the PhD degree from the Department of Computer Science and Engineering at the University of Notre Dame, Notre Dame, Indiana in 2005. His current research interests are the use of a multimodel framework to simulate avian limb growth, and algorithms for predicting protein-protein interactions.



applications of information theory in molecular biology. He is a student member of the IEEE.

**Faruck Morcos** received the BS degree in electronics and communications engineering from ITESM, Monterrey, Mexico, in 2001 and the MS degree in communications engineering from the Technische Universität München, Munich, Germany, in 2004. He is currently a PhD student in the Department of Computer Science and Engineering at the University of Notre Dame, Indiana. His current research interests include systems biology, protein networks, and



**Simon P. Kanaan** is a graduate of the Department of Computer Science and Engineering at the University of Notre Dame, Indiana. He is currently employed by Accenture in Chicago.



**Stefan Wuchty** received the MS degree in chemistry and the PhD degree in theoretical biochemistry and bioinformatics, both from the University of Vienna, Austria. He is a fellow at the Northwestern Institute on Complexity at Northwestern University, Evanston, Illinois. His current research focuses on the investigation of networks in areas as diverse as molecular biology, sociology, and business.



**Danny Z. Chen** received the BS degrees in computer science and in mathematics from the University of San Francisco, California, in 1985, and the MS and PhD degrees in computer science from Purdue University, West Lafayette, Indiana, in 1988 and 1992, respectively. He has been on the faculty of the Department of Computer Science and Engineering at the University of Notre Dame, Indiana, since 1992 and is currently a professor. His main research

interests are in algorithm design, analysis, and implementation, computational geometry, parallel and distributed computing, computational medicine, data mining, robotics, and VLSI design. He has published more than 130 journal or conference papers in these areas. In 1996, Dr. Chen received the Faculty Early Career Development (CAREER) Award of the National Science Foundation. He is a senior member of the IEEE and a member of the IEEE Computer Society.



**Jesús A. Izaguirre** received the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 1999. He is an associate professor of computer science and engineering at the University of Notre Dame, Notre Dame, Indiana. His current research is on efficient methods in chemistry and biology, particularly molecular dynamics, Monte Carlo methods, cellular automata, and analysis of biological networks. He is also interested in the

portable implementation of high performance software for scientific computing. Dr. Izaguirre received a CAREER Award from the US National Science Foundation in 2001, and a BP Foundation Outstanding Teacher of the Year Award in 2005. He is a member of the IEEE and the IEEE Computer Society.

► For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).